Goals for Concept Representation in the GALEN project

AL Rector WA Nowlan Andrzej Glowinski
Medical Informatics Group, Department of Computer Science
University of Manchester, Manchester M13 9PL

ABSTRACT

The GALEN project aims to develop language independent concept representation systems as the foundations for the next generation of multilingual coding systems. Traditional coding schemes have reached the limits of what can be maintained and managed, and a shift to formal compositional systems is now essential. GALEN is developing one such scheme and an associated concept model, together with criteria for their evaluation. It should provide the flexibility required to cope with the diversity amongst medical applications, whilst ensuring the coherence necessary for integration and re-use of terminologies.

INTRODUCTION

GALEN is an European Community funded project involving ten partners and six subcontractors in seven countries. It aims to develop language independent concept representation systems as the foundations for the next generation of multilingual coding systems. The GALEN project is based on the belief that the major obstacle to large scale integrated applications of advanced informatics in medicine is the lack of a standard representation for the concepts which are the basic units of medical information. However, universal prescriptive standardisation of medical information systems is neither possible nor desirable. GALEN aims to support the flexibility required to cope with the diversity amongst medical applications, while ensuring the coherence necessary for integration and re-use of terminologies.

GALEN has two aims:

i) To develop and validate a 'Master Notation for medical terminology' and Coding Reference (CORE) Model formulated in that formalism.

ii) To evaluate the usefulness and effectiveness of the Master Notation and CORE Model through demonstrators and experiments.

and two corresponding hypotheses:

i) Possibility — That the proposed techniques are sufficient to develop a Master Notation and Coding Reference (CORE) Model of medical terminology which are language independent, formally sound, and which can be scaled up smoothly.

ii) Usefulness — That the knowledge in the CORE Model will provide a major part of the common background knowledge required to develop and integrate systems serving many different needs, including decision support, medical records, bibliographic retrieval and natural language processing.

BACKGROUND: GOALS AND LIMITATIONS OF MEDICAL TERMINOLOGIES

GALEN is proposing alternative approaches to medical terminologies. To justify this it is necessary to consider the background to traditional schemes and understand why changing requirements have exposed the fundamental limitations of those schemes.

Goals of Traditional Coding Schemes

For over one-hundred and thirty years the systematic collecting and recording of medical information has been based on the use of enumerated terminologies formulated as classifications, nomenclatures, and coding schemes of various kinds. Until relatively recently these schemes were used mainly for recording causes of death and gathering minimal diagnostic information for statistical and epidemiological purposes. For these purposes, these schemes have needed to support:

- a limited range of uses;
- a small number of summary descriptions;
- manual use and direct human interpretation;
- normative use in a single language and linguistic community;
- rare and controlled extensions.

In recent years there has been a proliferation in the uses for systematically recorded medical information — in computer based medical records, quality assurance, medical audit, medical knowledge-based systems, and bibliographic systems, to name but a few. This has resulted in fundamental changes in the requirements for terminologies which now need to support:

- multiple uses for information;
- large numbers of complex detailed descriptions *e.g.* clinical descriptions in medical records;
- automatic manipulation by large computer-based information systems;
- multi-lingual and multicultural use;

0195-4210/92/$5.00 © 1994 AMIA, Inc.

- constant and frequent extensions.

These requirements are the converse of the original assumptions that justified the use of traditional techniques of classification. The first response to these new requirements has been to extend the content and structure of traditional schemes. Several of the major schemes have increased in size and complexity by one or more orders of magnitude. A second approach has been taken by the UMLS project [1] in providing an extensive external cross reference to the existing schemes. SNOMED III has taken a major step forwards in providing additional axes and making the existing axes more homogeneous. However, even SNOMED-III relies on enumeration within each axis.

ANALYSIS OF THE OBSTACLES TO WIDER USE OF EXISTING SCHEMES

The Limits of Unaided Human Effort

Traditional schemes are reaching the limits of what can be maintained and managed. The problem is not simply that the systems are large. Medicine is large and complex; hence anything which represents medical concepts will be, in some respect, large and complex. The problem is that traditional schemes rely solely on human effort to enumerate this complexity. Even the UMLS ultimately depends on coding systems which are maintained manually. There are several obstacles which face manually enumerated schemes.

The 'combinatorial explosion'

The first major problem with enumerative classification is that an apparently modest increases in the expressive power of a scheme can produce a combinatorial increase in the number of terms and relationships. Consider the case of burns which should be classified by site on the body, penetration, and whether thermal or chemical. The total number of terms for all possible combinations is on the order of:

(200 body sites +1) x (4 penetrations +1)
x (2 aetiologies +1) = 3015

Adding epidemiologically useful concepts such as whether the injury occurred at work or home, details of circumstances, and complications quickly brings to total to over 1,000,000. Worse, if users wish to query the system along several axes, each term might be classified in several different ways multiplying the total number of terms by another factor of three to five relations per term.

Mixing of concepts and relationships

The second major problem with enumerative classifications is that they mix different kinds of information without distinctions within the same structure. This is inevitable if the scheme is arbitrarily confined to a one or a few axes. While a suitable mix for any one application can often be achieved, the choices are almost always inappropriate for other uses, making re-use of the information difficult or impossible. Two types of mixing can be identified:

- Mixing of different kinds of concepts and different compositional principles in a single hierarchy — for example mixing of the reason for an action and the nature of the action in the same hierarchy. For example in recent work, the following hierarchy was suggested:

respiratory condition
    pleural effusion
        pleural effusion seen on radiograph
        pleural effusion diagnosed by percussion
        pleural effusion fluid

- Mixing of part-whole and generic relations, to take an example from SNOMED-II:

Bone
    Long Bone
    Periosteum
    Shaft

In each case the order in which the different criteria are applied and which combinations of characteristics are used in abstractions depends on the use to be made of the information. Making such decisions in advance restricts the usefulness of the information for other purposes.

Any attempt to deal with multiple axes or multiple classifications within a traditional framework aggravates the combinatorial explosion by increasing the number of relationships still further. While it is perhaps conceivable that unaided human effort could enumerate all of the terms required, it is inconceivable that it could identify, arrange and maintain all the required relationships.

Mixing of concepts and language

Not only have traditional systems mixed different conceptual issues, they have also tended to mix linguistic issues with the underlying concepts. Because traditional schemes rely on human interpretation, issues of the precise language used for concepts — of preferred terms, synonyms, and lexical variants — are intimately associated with and difficult to separate from the concept system itself. Such schemes may be translated, but the translation is nearly as great a task as the development of the original system.

Lack of formal structure.

Traditional systems can only be extended by reference back to their developers. Even if a concept is implied by the other concepts around it — a burn for a more detailed site or an infection by an unusual organism for example — it cannot be generated automatically. Furthermore, the new concept must be placed in the hierarchies according to the existing, often implicit, conventions— an error prone task, particularly with multiple classifications.

A major consideration is that changes should have no unanticipated side effects. Simple classifications contain little formal structure. The interpretation of the 'meaning' of codes is largely in their rubrics for interpretation by human users rather than in the model itself. Hence, there is no way for software to determine whether the implications of a proposed modification are 'safe'. Paradoxically, the goal of flexibility in use can only be achieved by formal rigidity in structure which allows verification of the consequences of change.

Unsuitability for automatic manipulation

A second consequence of the lack of formal structure is that the amount of computer-based manipulation which is possible, whether for decision support, user interface, or information retrieval is limited. User interfaces of systems using simple hierarchical systems have shown serious limitations. If computer systems are to manipulate concepts in ways which seem intuitive to human users, then much more of the structure of those concept systems must be modelled formally so as to be accessible to the computer.

GALEN'S APPROACH

Overall Goals in Developing the GRAIL Kernel

GALEN's approach to meeting these obstacles and requirements is to develop a fully compositional and generative system for modelling concepts — the GALEN Representation and Integration Language (GRAIL) Kernel. GRAIL uses two primary mechanisms

i) 'Particularisation' for composing concepts and coordinating hierarchies, *e.g.* Fractures which haveLocation Femurs.

ii) 'Sanctioning' for indicating which particularisation's are coherent with respect to any given model; e.g. Fractures sensibly haveLocation bones.

Particularisation, indicated by the key word which creates a new kind of an existing concept by adding a new criterion. Sanctioning is a single, unified mechanism analogous to type constraints in other languages for indicating which particularisation's are sensible and allowing redundant criteria to be recognised and removed. Roughly GRAIL can be seen as an 'extended Terminology Box', in the tradition of Brachman and Levesque's KRYPTON [2].

There are several goals for the GRAIL formalism:

i) To overcome the combinatorial explosion of terms in enumerative systems and the generation of nonsensical terms in partially compositional systems;

ii) To provide for the coordination of independent taxonomies and provide a strict separation of generalisation and part-whole relationships;

iii) To provide a clean separation between the concept model and linguistic mechanisms which interpret that model in order to allow the development of multilingual systems;

iv) To provide the formal structure required for automatic manipulation of concepts which appears intuitive to human users;

v) To be formally sound and produce models that are verifiable and contain no contradictions or ambiguities;

vi) To be computationally tractable, and give rise to models which are scalable;

vii) To be organisationally manageable and capable of being maintained with realistic human effort.

What must be represented

In order to gain a greater insight into what needs to be represented consider the following example — 'Pathological Fracture of the neck of the Femur caused by Osteoporosis'. Clearly this concept is made up of a number of different elementary concepts. Potentially it might be classified as being a kind of the concept of 'fractures of femur', the concept of 'pathological fractures', the concept of 'conditions caused by osteoporosis', or even 'fractures of the shaft of long bones'. It might even be considered whether 'pathological' is redundant, since a fracture caused by osteoporosis might be considered to be 'pathological' by definition. The question is what 'knowledge' needs to be represented in order to produce a model of our understanding of this concept? A short list must include:

i) *What basic kinds of things exist:* Fractures exist and are conditions and femurs exist and are bones. In GRAIL:

Conditions subsume Fractures.
Bones subsume Femurs.

ii) *What is 'sensible':* Fractures may occur to the femur — or more generally to bones; the femur has a part (division) known as the 'neck'; the neck of the femur may fracture — or more generally, divisions of bones may fracture. In GRAIL:

```
Conditions grammatically haveLocation
                                    BodyParts.
Fractures sensibly haveLocation Bones.
Femurs sensiblyAndNecessarily haveDivision
                                          Neck.
```

iii) *How do partitive relations and descriptions fit together:* Fractures of the neck of the femur are fractures of the femur — or more generally, fractures of divisions of bones are fractures of the corresponding bones, or more generally still that conditions of divisions of things are conditions of the corresponding thing. In GRAIL:

```
haveLocation refinedAlong hasDivision
```

iv) *What are the definitions:* What is it we mean by 'pathological' in 'pathological fractures'? (There are at least three possible definitions: a) fractures in which there is a cause other than external trauma; b) fractures in which external trauma is definitely not one of the causes; c) fractures for which the role of external trauma is not specified. If in GRAIL we adopt the positive meaning for 'pathological' i.e. having a physiological cause then the definition:

```
(Fractures which haveCause-
    PathologicalConditions) name
            PathologicalFractures.
```

Given this knowledge we can now generate the composite particularisation:

```
Fractures which
    <haveCause-Osteoporosis
    haveLocation-(Neck which areDivisionsOf-
            (Femurs which  haveLaterality-right))>
```

and classify it correctly as being subsumed by:

```
Fractures which haveCause Osteoporosis,
Fractures which haveLocation Femurs,
etc.
```

More importantly, most of this knowledge can be reused to generate and classify a large number of related concepts, for example a pathological fracture of the neck of the humerus. At the same time, the sanctioning mechanism prevents generation of 'nonsensical' compositions such as Fractures which haveLocation Blood. A key principle of GALEN is that the total number of facts required to model any large terminology should be much less than the total number of 'sensible' compositions which can be generated and classified using those facts.

Objectives for GRAIL Models—Representation of Concept Systems and Formal Properties of Models

Based on examples such as the above the preliminary experience with the GRAIL Kernel and its predecessor SMK [3][4], eight aspects of our intuitive understanding of concept systems have been identified which require modelled:

i) The atomic concepts in the system;

ii) The generic relation which organises the atomic concepts into kinds;

iii) The rules for deciding which composite concepts are sensible;

iv) The essential characteristics of concepts, both atomic and composite;

v) The *formal* classification of concepts into 'kinds' based on their essential characteristics;

vi) The *naming* of composite concepts, or conversely the *definition* of concept names in terms of composite concepts.

vii) The interaction of part-whole relationships with the formal classification into 'kinds'.

viii) The equivalence of composite concepts involving tautologies and trivial variants

ISSUES

Language independent concept models - is there a shared medical model?

The assumption underlying the increasing internationalisation of medicine is that there is a shared model of medical science and medical care which transcends local idiosyncrasies of language and usage. It remains a matter of controversy as to how far such a language-independent model of medical concepts is possible. The first step in resolving this controversy is to attempt to develop formal systems which distinguish clearly between the concepts represented and the linguistic terms and mechanisms used to refer to those concepts.

Is a formal system possible?

Can a system be built which can be scaled up to realistic systems. We believe the answer to this question is yes and have evidence for this position [3][4], but it remains one of GALEN's fundamental question. Three different types of questions concerning scaling have been identified:

i) Formal worst-case scaling properties of the formalism. Any suitable formalism will be combinatorially explosive in the worst case, since it is based on combinations of criteria. The issue is to identify the parameters of the model which determine the formal scaling behaviour.

ii) Empirical scaling behaviour with medical data. The models which are being developed relate to practical cases which are sparse and whose behaviour is a matter for experiment. The goal is a formalism which is tractable when modelling real medical concept systems, not a guarantee that *any* large model is tractable. [5].

iii) The scaling in human effort required to build the models. The goal is to reduce the human effort required to develop the models. However, formal models introduce difficulties of their own. Their rigidity and formal structure imposes disciplines which are difficult for unaided users to control.

Would a formal system meet the requirements?

Cimino [6] lists seven characteristics required for a controlled medical vocabulary — what we have here called a 'terminology'.

i) Domain completeness

ii) Unambiguousness

iii) Non-redundancy

iv) Support for synonymy

v) Multiple classification

vi) Consistency of views

vii) Support for explicit relationships other than hierarchical relations.

The modelling approach which GALEN advocates will normally provide completeness in the sense that once sanctioning links are established, all possible composite statements are automatically sanctioned. There can never be arbitrary lacunae in the model, although there can of course be areas which are simply missing. GALEN has placed great emphasis on establishing the transformations required to reduce composite concepts to an unambiguous canonical form which excludes redundancy. GRAIL also supports multiple 'interpretations' of concepts, although, in general, problems of 'synonyms' are deliberately separated from the concept model and dealt with in a separate 'Multilingual Module'. Multiple classification and consistency of views are natural outcomes of the compositional modelling approach, and explicit modelling of additional relations is likewise intrinsic to the approach.

More generally, the question of whether the system meets requirements can only be tested through its use in applications. Hence applications form a major part of the GALEN project.

STATUS OF PROJECT

The first version of the 'master notation'— the GALEN Representation and Integration Language (GRAIL) Kernel, version 1 and the associated software, the terminology engine, have now been released along with the first portions of the COding REference (CORE) Model [7]. These are to undergo evaluation and be used in experiments during the summer in order to provide revised requirements for the second version in 1994. Early material has also been used at the recent CANON group workshop on medical concept modelling.

Additional information on the GALEN project and the deliverables can be obtained from the project coordinator at the above address.

REFERENCES

1] Humphreys BL and Lindberg DAB (1992). The Unified Medical Language System PROJECT: A distributed experiment in improving access to biomedical information. *in* KC Lun, P Degoulet, TEPiemme, O Rienhoff (eds). Proceedings of MEDINFO-92. Amsterdam, North Holland. pp 1496-1500

[2] Brachman RJ, Fikes RE and Levessque HJ (1985) An essential hybrid reasoning system *in Proceedings of IJCAI-85.* Morgan Kaufman. pp 532-539.

[3] Nowlan WA, Rector AL (1991) Medical knowledge representation and predictive data entry. In Stefanelli S, Hasman A, Fieschi M, Talmon J (eds) *AIME91, Lecture Notes in Medical Informatics no 44,* Springer–Verlag, Berlin 1991, 105–116

[4] Rector AL, Nowlan WA and Kay S.(1991). Conceptual Knowledge: The Core of Medical Information Systems. *in* KC Lun, P Degoulet, TE Pierre, O Rienhoff (eds) *MEDINFO 92,* North-Holland 1992, pp 1420-1426.

[5] Doyle J, Patil RS (1989) Two dogmas of knowledge representation: language restrictions, taxonomic classification, and the utility of representation services. Massachusetts Institute of Technology, MIT Report MIT/LCS/TM-387.b

[6] Cimino JJ, Hripcsak G, Johnson SB and Clayton PD (1989). Designing an introspective, multipurpose, controlled medical vocabulary. *in* LC Kingsland (ed). *Proceedings of the thirteenth annual Symposium on Computer Applications in Medical Care,* Washington, DC. IEEE Computer Society Press. pp 513-517.

[7] GALEN: The GRAIL Kernel, Version 1. GALEN Deliverable 6. The GALEN Consortium. Available from the authors.

Acknowledgements: This research supported in part by project A2012 under the EC initiative for Advanced Informatics in Medicine (AIM). The members of the GALEN consortium are: University of Manchester (Coordinator), Hôpital Cantonal Universitaire de Genève, Consiglia Nazionale delle Ricerche (Italy), Conser Sistemi Avanzati (Italy), The Finnish Hospital League, the Finnish Technical Research Centre, GSF-Medis Institut, (Germany), Hewlett-Packard Computers UK Ltd, Katholieke Universiteit Nijmegen (Holland), University of Linköping (Sweden), University of Liverpool.